Model Development Using Generalized Linear Models

September 11, 2023

Outline

- Basics of Supervised Learning
- Generalized Linear Model Overview and Model Fitting
- Variable Selection and Variable Transformation/Encoding
- Multicollinearity: Checking and Mitigation
- Model Training and Performance Assessment:
  Continuous and Binary Responses

Structured data

- Structured or tabular data with response/label
  - n observations
  - p covariates
- Examples of structured financial data
  - Fanny Mae data on Single Family home loan performance data
    - Performance of loan over time
    - Stacked snapshots of performance → structured data
    - Example Response : loan defaults in 12 months
    - Covariates : Property type, credit score, loan to value, price index, etc… See full list.
  - Credit risk assessment data
    - Response – approval of line of credit
    - Covariates – Mortgage, Balance, Amount Past Due, Delinquency Status, Credit Inquiry, etc…
    - Simulated internal data

Obs id	X_1	X_2	…	X_p	y
1					
2					
…					
n					

Data Pre-processing – structured data

Outlier detection

- Univariate methods
  - Quantile based outlier
  - Z-scores (normality assumptions). Sensitive to presence of outliers
- Multivariate methods(anomaly detection):
  - Isolation forest (Liu, Ting and Zhou, 2009)
  - Local outlier factor (Breunig et al. 2000)
  - DBSCAN (Ester et al. 1996)
- Treatment:
  - Data transformation (ex: log scale)
  - Clipping

Imputation

- Missing at random
  - Mean, median, mode, random sampling
- Missing conditionally at random
  - mean/median/mode/random sampling conditioned on class label
  - Model based approach
  - Binning approach
    - Bin non missing data of variable
    - Compare response of missing value to mean response of buckets and assign bucket
- Add indicator of missing data

Basics of supervised learning

Supervised learning and Loss functions

$$y = f(x) + noise$$

Goal:

Prediction

Estimation of $f(x)$

Given covariate information x, find $\hat{y} = \hat{f}(x)$ that minimizes some loss/cost.

Loss functions are designed specific to the learning task. Common loss functions for regression/classification

Continuous response

- Mean square loss : $\frac{1}{n}\sum_i(y_i - \hat{y}_i)^2$.
- MAE
- Huber Loss

Not sensitive to outliers

Binary Response

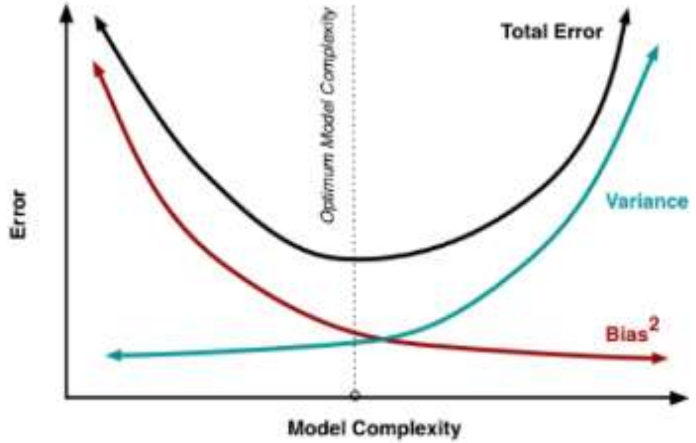
- Logloss $-\frac{1}{n}\sum_i[y_i \log\hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i)]$

Multinomial response

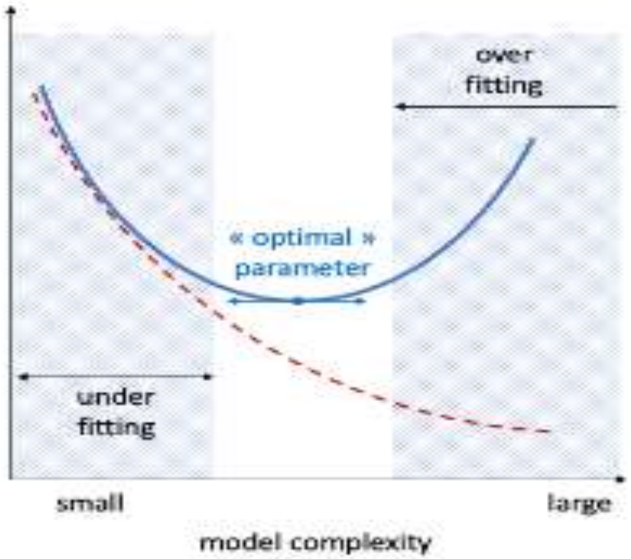
- Cross-entropy $-\frac{1}{n}\sum_i\sum_j\left[y_i^{(j)}\log\hat{y}_i^{(j)}\right]$

Bias variance tradeoff and over-fitting

- All learning algorithms come with hyper-parameters (HP) which controls the complexity of the algorithm. Examples:
  - High dimension in linear/logistic regression : Large number of predictors
  - Degree of polynomial in polynomial regression
  - Number of knots in splines
  - Tree based ensemble learners : depth, learning rate, minimum samples per leaf
  - Neural Network : number of layers, layer sizes,



- Complexity of a model is related to it's bias-variance trade-off
  - Underlying model : $y = f_\star(x) + \epsilon$
  - $E\left(\hat{f} - f_\star\right)^2 = Var\left(\hat{f}\right) + Bias\left(E\left(\hat{f}\right), f_\star\right)^2$
  - Very complex (over-fitted) model – low bias, high variance
  - Simple (under-fitted) model – high bias, low variance



-- training error — validation error

Generalized Linear Model Overview

Brief over-view of statistical approaches

$$y = f(x) + noise$$

- Parametric approaches
  - $f(x)$ is modelled as an explicit function of x and some parameter β
  - β is finite dimensional
  - Examples
    - Linear regression, logistic regression, polynomial regression
- Semi-parametric approaches
  - Underlying parameter has a fixed dimensional component and a component that can grow with the number of training observations
  - Examples
    - B-Splines, cox-proportional hazard models
- Non-parametric approaches
  - Underlying parameters can grow with the number of training observations
  - Classification and Regression trees
  - Ensemble models
  - Kernel regression

Linear Regression

- A model often used to explore the relationship between a continuous response and a set of covariates.
- Let *{y_i}* be continuous responses and *{x_i}* be a set of *p* predictors.
- The linear regression model is:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i} + \epsilon_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i$$

- The goal is to estimate the regression coefficients using data, predict the response, select appropriate variables, etc.
- Predictions may be made using the estimated coefficients:

$$\hat{y}_i = E[Y|X] = \widehat{\beta_0} + \widehat{\beta_1} x_{1,i} + \cdots + \widehat{\beta_p}\, x_{p,i}$$

Statistical Inference for Linear Models

- With some additional assumptions, it is possible to conduct statistical inference on the coefficients and the model, such as:
  - Confidence intervals and hypothesis tests about each coefficient
  - Overall tests of model fit
  - Confidence intervals for the mean response and prediction intervals for observations
- Additional modeling assumptions must be assessed for the inference to be valid; variants can be used in some cases.
- Such analysis is not necessary if prediction is the only goal, but can be useful in understanding the data and the model.

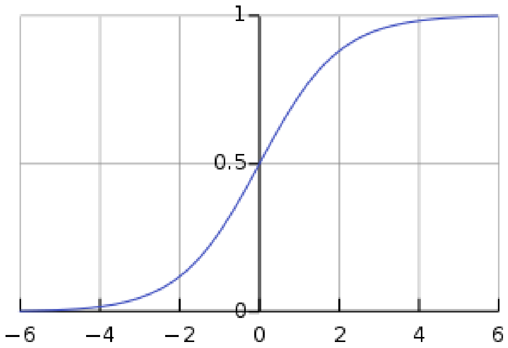
Generalized Linear Models

- In many applications, the response variable can take a variety of types, such as binary, categorical, counts, ordered categorical.
- The linear regression model is not always appropriate for modeling these types of responses.
- The Generalized Linear Model (GLM) framework is a universal approach to handling such types of data.

<u>In the GLM framework:</u>

- A distribution for the response Y is appropriately chosen
- A function of the conditional mean of the response is modeled as a linear combination of the predictors:
    - $g(E[Y|X]) = g(\mu) = \boldsymbol{X\beta} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$
- where $g(\cdot)$ is called the link function, and relates the linear combination of predictors to the mean of the distribution of y
- Note that $E[Y|X] = g^{-1}(\boldsymbol{X\beta})$

Common Uses of the GLM Framework



For logistic regression:

$$g^{-1}(z) = \frac{e^z}{1+e^z}$$

- Common examples of GLMs:
  - In linear regression: $y \sim N(\mu, \sigma^2)$ with $g(\mu) = \mu = X\beta$ (the link function is the identity)
  - In logistic regression, for binary responses, we have:

    $$y \sim Bernoulli(\pi) \text{ with } g(\pi) = log\left(\frac{\pi}{1-\pi}\right) = X\beta$$
  - In log-linear regression, for count responses, we have:

    $$y \sim Poisson(\mu), \text{ with } g(\mu) = \log(\mu) = X\beta$$

- The **inverse** of the link function, also called the mean function, is useful for making predictions

- Note that there are several other examples in this framework, and in some cases a variance component must be estimated in addition to the mean. Statistical inference theory is well-established for these models as well.

Estimation of GLM Models

- Training a model in this family means to estimate the values of the parameters (coefficients).
- To estimate the parameters of a GLM, we minimize an appropriate loss function (or find the parameters that maximize the likelihood function):
  - **Continuous response:** Mean square loss : $\frac{1}{n}\sum_i (y_i - \hat{y}_i)^2$.
  - **Binary Response:** Logloss $-\frac{1}{n}\sum_i [y_i \mathrm{log} \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)]$
- Sometimes, this can have a closed form solution, such as in the linear regression case (vector version):

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

- Other times, as in logistic regression, there is no closed-form solution. In this case, we can use an iterative algorithm, such as Newton-Raphson to update the estimated parameters until convergence.
- In practice, we typically rely on software to fit such common models.

Regularization

- Regularization : Penalty to loss function for increase in model complexity

$$Regularized\ loss = Loss(y, x, \beta) + \lambda \times Penalty(complexity)$$

  - Penalty parameter λ is treated as Hyper-Parameter (HP) in model.
  - Tune λ to obtain optimal model that minimizes hold-out validation error
  - **Objective:** force some parameters to be small or 0 thereby reducing model complexity without compromising on performance too much.
  - **Result :** Reduce over-fitting and generalizable model performance

- Common Types of Regularized Regression:
  - Ridge Regression or L2 penalty: Minimizes $Loss + \lambda \sum_j \beta_j^2$
    - Coefficients moved towards 0 but not exact sparsity
    - Effective in presence of multi-collinearity
  - LASSO Regression or L1 penalty: Minimizes $Loss + \lambda \sum_j |\beta_j|$
    - Coefficients shrunk to zero; exact sparsity
  - Elastic Net: Includes both L2 and L1 penalties.

Variable Transformation/Encoding and Variable Selection

Variable Transformation/Encoding

- Sometimes necessary to transform or encode variables before including them in a model
- **Response:** Handle outliers, stabilize variance, or otherwise make the response more suitable for the model. Typical transformations include:
  - Log transformation for positive, long-tailed response, e.g. time-to-event data:

$$g(y) = \log(y)$$

  - Box-Cox family of power transformations:

$$g_\lambda(y) = \frac{y^\lambda - 1}{\lambda}$$

- **Categorical Predictors**:
  - Typically, *k* categories are encoded as *k-1* **indicator variables** (or dummy variables) that take the value of 1 if the observation is in that category, and zero otherwise.
  - Also called "one-hot encoding". In ML, may include all **k** indicators.
  - ML models offer categorical embedding and others, depending on the ML technique. Often useful for large numbers of categories.

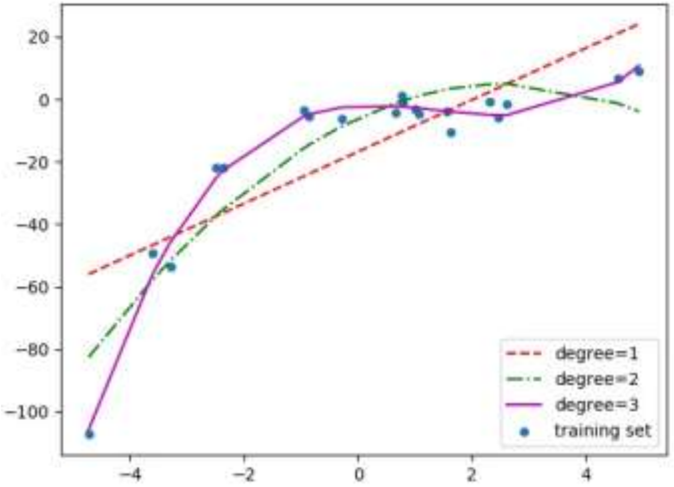
Variable Transformation/Encoding

- **Numerical Predictors:** Interaction Terms for effect of combinations of variables
  - Interaction Terms: New terms in that include the impact of two or more other predictors.
  - Traditional approach: multiplicative interactions, but can be generalized:

$$\hat{\boldsymbol{Y}} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{1,2} X_1 X_2$$

  - Examples: Non-linear with continuous response
    - Polynomial regression X^2, X^3
    - Linear model with interactions $y = \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \beta_3 \mathrm{X}_3 + \beta_{12} X_1 X_2 + \beta_{123} X_1 X_2 X_3 + \epsilon$
    - Polynomial with interactions $y = \beta_0 + \beta_1 \mathrm{X}_1^2 + \beta_2 \mathrm{X}_2^3 + \beta_3 \mathrm{X}_3 + \beta_{12} X_1 X_2 + \beta_{123} X_1 X_2 X_3 + \epsilon$

- **Numerical Predictors:** Non-linear relationships
  - Pre-defined functional transformations, e.g. $x^2, \log(x)\,, \sqrt{x}$, etc
  - Binning based on value, creating indicator variables for each bin.
  - Flexible/general approaches, e.g. splines (next slide)



Splines– Flexible Non-linear transformations

- Piece-wise functions used to model effects of each variable smoothly. Examples:
- Linear splines (derivatives not continuous at knots)

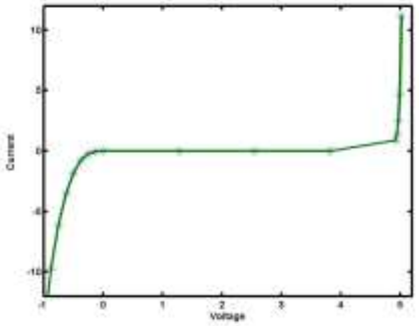
$$f(x) = \beta_0 + \beta_1 x + \Sigma_{k=1}^{K} b_k (x - \xi_k)_+ \quad , \xi_1 < \xi_2 < \cdots \xi_K$$

- Truncated power basis splines of order q (q-1 derivatives continuous)

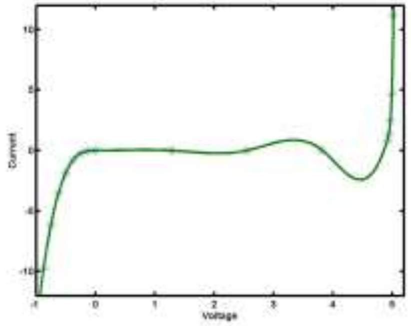
$$f(x) = \beta_0 + \beta_1 x + \cdots \beta_q x^q + \Sigma_{k=1}^{K} b_k (x - \xi_k)^q{}_+ \quad , \xi_1 < \xi_2 < \cdots \xi_K$$

- B-Splines
  - $f(x) = \Sigma_{k=1}^{K} b_k B_{kd}(x)$
  - $B_{kd}(x) : k_{th}$ spline of degree d
  - Defined through recursive relations with
  - $B_{k0}(x) = 1 \; if \; \xi_k \leq x < \xi_{k+1}$
  - $B_{k(j+1)}(x) = \alpha_{k(j+1)}(x) B_{kj}(x) + (1 - \alpha_{(k+1)(j+1)}(x)) \, B_{(k+1)j}(x)$
  - Efficiency in computation
  - Any spline function can be expressed as linear combination of b-splines

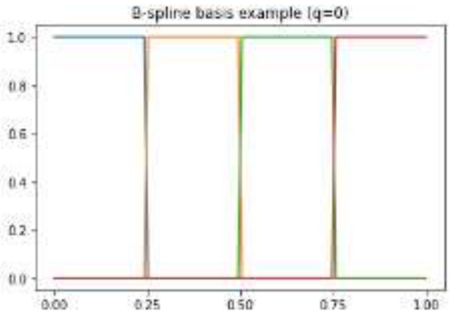
linear

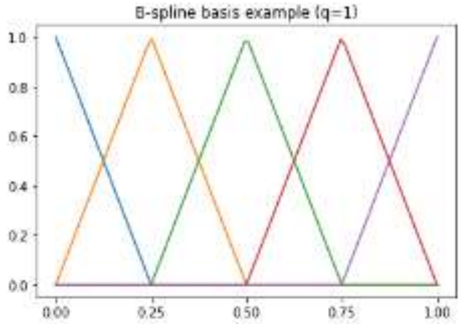


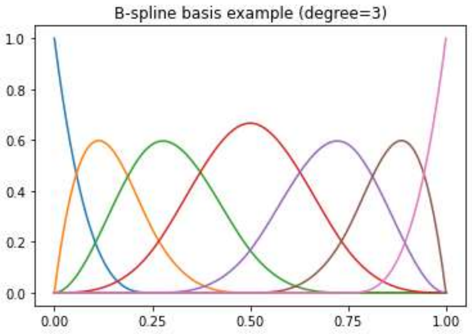
cubic



B-splines







GAM and splines

Generalized Additive Model (GAM) :

$$E(Y|X) or\ logodds(p) = f(X) = \alpha + f_1(X_1) + f_2(X_2) + \ \ldots + f_p(X_p)$$

$f_j(X_j)$: unspecified smooth non-parametric functions

- Each $f_j(X_j)$can be estimated through some spline function. Example:

$$\hat{f}\ (X) = \ \Sigma_{j=1}^{p} \Sigma_{k=1}^{K(j)} \beta_{jk} B_{jkd}(X_j)$$

- Number of knots and order define complexity of model

Variable Selection: Motivation

- Many variables may be available to improve model fitting.
- However, one needs to select important or significant variables to enhance predictability, interpretability, generalizability, and avoid over-fitting.

- Principle of **Parsimony**: Generally, prefer a simpler model over a more complex model if both fit the data well.

Variable Selection Criteria

- Most variable selection criteria are based on residual sum of squares:

$$RSS = \sum_i (y_i - \hat{y}_i)^2$$

- Performance-based variations include:
  - Mean Squared Error, $MSE = \frac{RSS}{n}$
  - R-squared, $R^2 = 1 - \frac{RSS}{\sum_i (y_i - \bar{y})^2}$
- Variations balancing performance and complexity (for linear models- generalizations are available)
  - Adjusted R-squared, $R^2_{adj} = 1 - \frac{n-1}{n-p}(1 - R^2)$
  - Mallow's C_p statistic, $C_p = \frac{RSS}{\sigma^2} - n + 2p$
  - Akaike's Information criterion, $AIC = RSS + 2p\sigma^2$
  - Bayesian Information Criterion, $BIC = RSS + \log(n)\, p\, \sigma^2$

Classical Variable Selection

- **Best Subset Selection**: Exhaustively searches all possible subsets of predictors, and calculates the criterion/score for each subset.
- **Forward Selection**: Starts with a model with no predictors, then at each step, search all remaining candidate predictors for the one that gives the greatest improvement in the criterion, and include that predictor if the improvement in score exceeds the pre-specified threshold. Stop when no additional predictors improve the fitting beyond the threshold.
- **Backward Elimination**: Starts with a model with all candidate predictors, then eliminates the least important predictor if the reduction is greater than a pre-specified threshold, until no other predictors can be discarded.
- **Stepwise Regression**: A combination procedure of forward selection and backwards elimination

- Both forward and backward elimination are problematic when variables are correlated, because in importance of variables in the model can change as other variables are added or removed. For example, in forward selection, an already-added variable can become unimportant if a correlated variable is added.

Variable Selection (Continued)

- Other algorithmic selection approaches:
  - **Least Absolute Shrinkage and Selection Operator (LASSO)**: Penalized Regression approach that shrinks coefficients towards zero; may be used to eliminate variables with a coefficient of zero.
  - **Least Angle Regression (LARS)**: Algorithm for fitting regression, starting with all coefficients equal to zero. Selectively increase those most correlated with the response.
- Other approaches:
  - Causal Variable Selection: Borrow techniques from causal discovery to select a set of features such that the response is independent of the other candidate features, conditional on this set.
  - Strength of relationship: Choose variables that show strong relationships with the response
  - Feature Importance from machine learning algorithms to identify important variables
- **Note:** Regardless of the method, consider the purpose of the model, and the reasonableness of the selected variables in the context of the problem being modeled!

Multicollinearity: Checking and Mitigation

Multicollinearity

- Definition: Occurs when one or more predictors are (nearly) linear combinations of other predictors
  - Perfect or exact multicollinearity occurs when one predictor is an exact linear combination of other variables, e.g. a dataset of payroll information including each employees withholdings, net, and gross pay.
  - Near multicollinearity occurs when two or more predictors are highly correlated with each other. (e.g. different monthly unemployment calculations)
- Diagnostics:
  - Variable Inflation Factor (VIF)
  - Examining the correlation matrix of the predictors
- Consequences:
  - Numerical difficulty or instability when fitting the model.
  - If interested in statistical inference, standard errors of the coefficient estimates will be inflated; more uncertainty in the estimate.
  - Even if not interested in statistical inference, the model training may allocate explanatory power to the correlated features in undesirable ways

Multicollinearity: Solutions

- Use *ridge regression* to stabilize model estimation, if that is a problem.
- Consider variable selection techniques to choose the most important variable from a set of correlated predictors.
- Alternatively, engineer a new variable from the sets of correlated features. This can be done via either a dimension reduction technique (e.g. PCA) or business insight.

- Note: If prediction of the response is the only consideration, this may not be an issue; however, if understanding the model is important, consider the impact of (near) multicollinearity.

Model Training and Performance Assessment: Continuous and Binary Responses

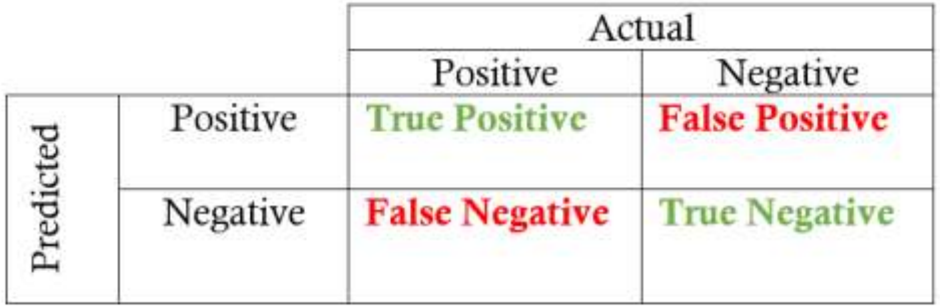
Performance Assessment for Continuous Response Models

- Different metrics are often used to give different views on the model performance.
- Some common metrics are:
  - Mean Squared Error (MSE): $\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$
  - Root Mean Squared Error (RMSE): $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$
  - Mean Absolute Error: $\frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$
  - R$^{2:}$ $1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}$

Performance Assessment for Binary Response Models

- Score-Based Metrics: Useful when interested in a predicted score rather than a class
  - Log-likelihood, or -2 log likelihood: sensitive, but scale is data-dependent
  - AUC- “Area Under Curve”, specifically the Receiver Operating Characteristic Curve
    - Scale is independent of data

- Binary Prediction Based Metrics: Useful if a class prediction is the goal
  - Accuracy
  - Confusion Matrix Based Metrics, like:
    - Recall = True Positives/(True Positives + False Negatives)
    - Precision = True Positives/(True Positives + False Positives)
    - Etc….



		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Next Steps

- Questions
- Examples